BIG DATA EUROPE

Support Action

Big Data Europe – Empowering Communities with Data Technologies

Project Number: 644564   Start Date of Project: 01/01/2015   Duration: 36 months

Deliverable 6.5 Pilot Deployment II

Dissemination Level	Public
Due Date of Deliverable	30 April, 2017
Actual Submission Date	8 June, 2017
Work Package	WP6, Real-life Deployment & User Evaluation
Task	T 6.2
Type	Other
Approval Status	Pending
Version	1.0, Final
Number of Pages	23
Filename	6.5 Pilot Deployment II.pdf

Abstract: This document is the reference to D6.5 classified as status "Other" which in this case means the implementation and deployment efforts on the pilots for the 7 Societal Challenges. The source code, scripts, recorded demonstrations and manuals for deploying the pilots are the core of this deliverable. This report serves as an index for these various online sources on the first cycle of pilots.

History

Version	Date	Reason	Revised by
0.1	18/05/2017	Final Draft	Ronald Siebes
0.2	25/05/2017	Reviewed version	Aad Versteden
1.0	16/08/2016	Final draft corrected according to review comments	Ronald Siebes

Author List

Organisation	Name	Contact Information
Open PHACTS foundation	Bryn Williams-Jones	bryn@openphactsfoundation.org

VU University Amsterdam	Victor de Boer	v.de.boer@vu.nl
VU University Amsterdam	Ronald Siebes	rm.siebes@few.vu.nl
NCSR-D	S. Konstantopoulos, A. Charalambidis, I. Mouchakis, G. Stavrinos	konstant@iit.demokritos.gr
TenForce	Aad Versteden, Erika Pauwels	aad.versteden@tenforce.com

Executive Summary

This document serves as an index for references to the implementation and deployment efforts that realized the second and partially the third cycle of pilots for the seven Societal Challenges. The source code, scripts, recorded demonstrations and manuals for deploying the pilots are the core of this deliverable. The good collaboration between the technical and domain partners combined with deliverance of a working infrastructure, generic Big-Data components, user interfaces and documentation   according to internal deadlines resulted in a versatile but coherent set of demonstrators. Various webinars, hangouts and presentations during the last year are recorded and serve as excellent reference material in this document together with the public code repositories where the more technically skilled reader can get more insight into the details.

The pilots demonstrate how relevant large-scale datasets or data-streams for the respective seven SC communities can be processed by the BigDataEurope infrastructure and provide novel insights that are promised by the Big Data community.

This document refers to online demonstrations, presentations, instructions and code-bases that form the content of D6.5.

Abbreviations and Acronyms

LOD	Linked Open Data
SC	Societal Challenge
BDE	Big Data Europe

Table of Contents

1.　Introduction

Implementation and instantiation are the two main aspects of the pilot deployment. These consist of having a code-base, a CPU/GPU cluster to test and deploy the code, creating modules by the various partners either from scratch or, in most cases modifying existing code into Docker components. Close collaboration with the technical partners (TenForce, SWC, NCSR-D and FhG) that are responsible for creating the BDE infrastructure and the generic components is essential. It is the role of WP6 to facilitate the communication between the technical partners and the domain partners (those who have the specific knowledge for each Societal Challenge) and guide the pilot implementation process. This report serves as an index referencing the current implementation and deployment work for the second and partially the third cycle of pilots for the seven Societal Challenges which are available online and which are part of this deliverable.

The best way to get a concise technical overview of the development and deployment status of the Big Data Europe platform:

https://www.big-data-europe.eu/webinar/#launch

The best way to get a written overview of each pilot and the relation to the generic BDE components and the infrastructure is reading deliverable D5.5.

The best way to dive into the source code and deployment scripts is to browse through the sub-repositories starting with 'pilot-SCxxx' on the main BDE GitHub repository:

https://github.com/big-data-europe

2. Planning

Figure 3 shows the initial planning for the implementation and deployment of the pilot cycles as described in deliverable D6.1 and D6.4.
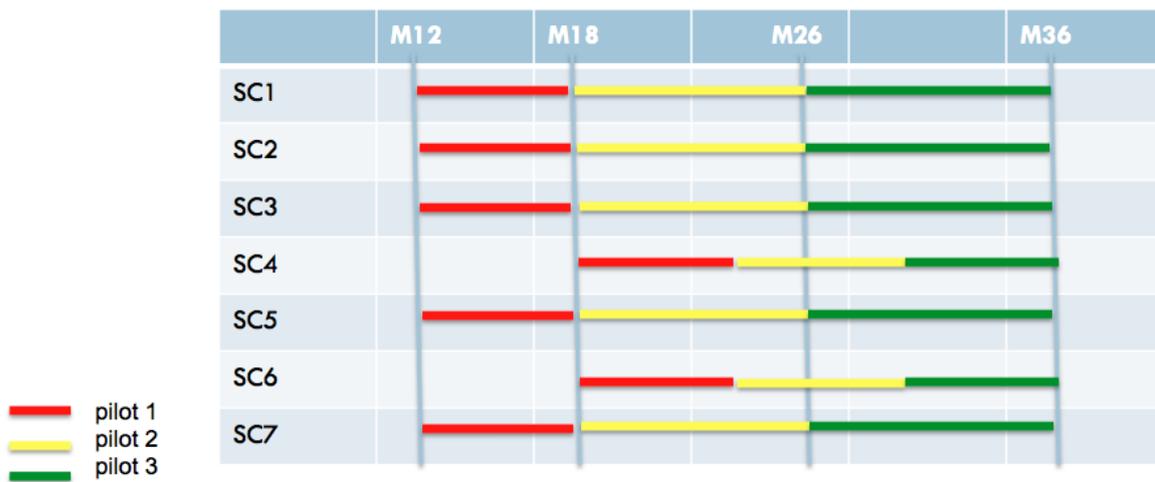

Figure 1: pilot planning

Currently the development and deployment is progressing according to schedule and we are entering the evaluation phase (deliverable D6.3).

3. SC 1 - Health - pilot

The pilot is carried out by OpenPHACTS and VUA in the frame of SC1 *Health, Demographic Change and Wellbeing*. NCSR-D (extending 4Store) and the University of Bonn (deploying the SANSA stack) provide the technical effort on testing various distributed RDF storage solutions.

The pilot is described in section 2 from deliverable D5.4.

The pilot implements the workflow of reproducing the functionality of an existing data integration and processing system (the Open PHACTS Discovery Platform) on BDI.

The most important challenge for this pilot is to replace the commercial cluster RDF[1] store, Virtuoso, with an open source variant. There is currently only one promising candidate, 4Store[2], which which has received virtually no updates in the past year. Therefore the VUA, OpenPHACTS and NCSR-D have decided to adopt the code and improve it in such a way that it can serve as a generic component on the BDE infrastructure. This component will not only serve SC1 but also SC2 (Agriculture) and SC6 (Social Sciences) because the *veracity* and

---

[1] http://virtuoso.openlinksw.com/features-comparison-matrix/

[2] https://github.com/garlik/4store

variety aspects also apply to them. Linked-Data is the likely candidate to deal with this for which a distributed RDF datastore is required when dealing with Big Data. The ongoing implementation work on the 4Store BDE docker component can be followed here: https://github.com/big-data-europe/docker-4store

The first test results were demonstrated at the FLINK workshop part of the Int. Semantic Web Conference last October in Kobe, Japan [3]

SC1 - Health - Pilot - Cycle 2 references
Presentations
● https://www.big-data-europe.eu/bde-at-ehealth-week-2017/   ● https://www.big-data-europe.eu/bde-sc1-third-hangout-open-phacts-re-engineered-with-big-data-europe/   ● https://www.big-data-europe.eu/hands-on-bde-health-at-ict-open-2017/   ● https://www.big-data-europe.eu/a-look-back-at-the-2nd-workshop-on-big-data-in-health-demographic-change-and-wellbeing/   ● https://www.big-data-europe.eu/benchmarking-semantic-query-processing-from-real-world-pharmacological-workflows/   ● https://www.big-data-europe.eu/the-open-phacts-pilot-second-hangout-for-the-health-societal-challenge/   ● https://www.big-data-europe.eu/installing-and-running-the-

---

[3] http://project-hobbit.eu/events/blink-2016/

	first-health-sc1-pilot/ ● https://www.big-data-europe.eu/event/webinar-pilot/
Instructions, data, deployment scripts and source code	● https://github.com/big-data-europe/pilot-sc1-cycle1 ● https://github.com/openphacts/ops-docker ● https://docs.google.com/document/d/1gO4qI1hfM_TPt4_XKKc96bRUxDbKSYwbShIq7z8b1lY/edit

4. SC 2 - Agriculture - pilot

The pilot is carried out by AgroKnow, FAO, and SWC in the frame of SC2 *Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy.*

The pilot is described in section 3 of deliverable D5.5.

One of the key challenges for this pilot is the automatic data extraction from heterogeneous scientific publications and tables. SWC took the lead here and worked hard on PoolParty[4] code to extract data from viticulture publications, to develop queries based on real-life research questions and to deploy efficiently on one of the BDE RDF storage components which are

---

[4] https://www.poolparty.biz/

accessed via one of the graphical web interfaces which are still under development. The goal of this second cycle is to showcase the ability of scalable processing workflows to handle a variety of data types (beyond bibliographic data) relevant to Viticulture.

SC2 - Agriculture - Pilot - Cycle 2 references	
Presentations	● https://www.big-data-europe.eu/2nd-european-policy-perspectives-on-data-intensive-agriculture-food-joint-workshop-of-big-data-europe-e-rosa/ ● https://www.big-data-europe.eu/vitis-big-data-demonstrator-agricultural-university-of-athens/ ● https://www.big-data-europe.eu/six-challenges-for-agriculture/ ● https://www.big-data-europe.eu/internet-of-food-farm-iof2020/ ● https://www.big-data-europe.eu/joint-workshop-of-big-data-europe-and-godan-for-societal-challenge-food-and-agriculture/ ● https://www.big-data-europe.eu/big-data-europe-platform-what-for-whom-and-how/ ● https://www.big-data-europe.eu/the-open-phacts-pilot-second-hangout-for-the-health-societal-challenge/ ● https://www.big-data-europe.eu/event/webinar-pilot/

Instructions, deployment scripts and source code	● https://github.com/big-data-europe/pilot-sc2-cycle1 ● https://docs.google.com/document/d/1UtOfVs9O5whfwFa9SE-kvgwv6DCLfQckKu-k9wBdQ-c/edit

5. SC 3 - Energy - pilot

The pilot is carried out by CRES in the frame of SC3 *Secure, Clean and Efficient Energy.*

The pilot is described in section 4 of deliverable D5.5.

The following are the supported scenarios:

A. Monitoring of distributed systems with on-line modules. The online capabilities will be implemented with protocols that are widely used for sensor data, thus increasing the pilot applicability to other domains (transport, industrial applications etc)
B. Perform long term monitoring of gearbox or blades with acoustic emission and support the research for preventive maintenance applications. This research will be complementary to the results of the first cycle of the pilot.

C. Inclusion of Non Destructive Testing for WT blades with Acoustic Emission module (high volume, research oriented)
D. Inclusion of 'real time' component (simulate SCADA-like data, investigation for OPC UA)
E. Extending the functionalities of the web app
F. Adaptations for application to other (stakeholder) data

SC3 - Energy - Pilot - Cycle 2 references	
Presentations	- https://www.big-data-europe.eu/bde-at-wind-power-industry-big-data-and-iot-forum-amsterdam-30-3132017/ - https://www.big-data-europe.eu/bde-presentation-to-wind-europe-wind-resource-assessment-workshop-2017/ - https://www.big-data-europe.eu/presentation-of-bde-at-windeurope-thematic-workshop/ - https://www.big-data-europe.eu/2nd-sc3-workshop-full-report-released/ - https://www.big-data-europe.eu/bde-pilot-case-in-energy-system-monitoring-in-wind-energy-production-unit/ - https://www.big-data-europe.eu/big-data-europe-on-line-hangout-in-energy-domain/
Instructions, deployment scripts	https://github.com/big-data-europe/pilot-sc3-cycle1

and source code	

6. SC 4 - Transport - pilot

The pilot is carried out by FhG and CERTH in the frame of SC4 *Smart, Green and Integrated Transport*.

The pilot is described in section 5 of deliverable D5.5.

A fleet of 1200 taxis generate Floating Car Data containing GPS location, speed and orientation of each vehicle. This data is collected in real time by a local taxi association for dispatching purposes and transferred in real-time to CERTH-HIT. After filtering and processing the data, traffic speeds for various streets are calculated in real time and provided to the citizens through the Mobithess and TrafficThess info-mobility portals.

The aim of the SC4 pilot is to address the scalability and fault-tolerance issues of those services using open source Big Data frameworks and evaluate new algorithms for monitoring and forecasting.

An important implementation part by CERTH in this pilot is writing 'R' scripts[5] that consume JSON data from Bluetooth data and other data on positions of taxi's in order to map match the position of the vehicles on a map. The data comes from a JSON service, is consumed by Kafka, pre-processed via FLINK, and sent to 'R-server' which performs the calculations (e.g. traffic status on a certain street). For the pilot MongoDB with Elasticsearch is chosen as the datastore allowing efficient visualisation via geo-spatial queries combined with static OpenStreetMap data.

For this second cycle the focus lies on predicting congestions and matching+visualising the respective road segments from the raw sensor data.

SC4 - Transport - Pilot - Cycle 2 references	
Presentations	- https://www.big-data-europe.eu/sc4-traffic-status-evolution-during-a-snow-event-in-thessaloniki/ - https://www.big-data-europe.eu/big-data-and-transport-where-can-it-take-us-presentations-from-the-hangout-available/ - https://www.big-data-europe.eu/behind-the-scenes-of-the-bigdataeurope-transport-pilot-recap-of-our-hangout/ - https://www.big-data-europe.eu/more-big-data-less-traffic-congestion/ - https://www.big-data-europe.eu/big-data-for-transport-webinar-wrap-up-the-tech-the-business-and-the-policy/

---

[5] https://www.r-project.org/

Instructions, deployment scripts and source code	● https://github.com/big-data-europe/pilot-sc4-flink-kafka-consumer ● https://github.com/big-data-europe/pilot-sc4-kafka-producer ● https://github.com/big-data-europe/pilot-sc4-pipeline ● https://github.com/big-data-europe/pilot-sc4-mapmatcher ● https://github.com/big-data-europe/pilot-sc4-docker-r

7. SC 5 - Climate - pilot

The pilot is carried out by NCSR-D in the frame of SC5 *Climate Action, Environment, Resource Efficiency and Raw Materials.*

The pilot is described in section 6 of deliverable D5.5.

In the domain of nuclear disasters, the timely and reliable prediction of the release origin, as well as its consequences, offers added value to facilitate decision making and response actions. The pilot for this second cycle deals with the problem when no information is known about the release itself but the only available measurements are in the form of time series of the substance

level along with current and past weather conditions. The measurements are assumed to be derived from one or more of the monitoring stations located within Europe and the emergency is also assumed to occur within Europe. The usual computational approaches in these cases involve working backwards from the current atmospheric conditions in order to estimate the source location, a process known as inverse modeling. However, large computational times are usually required, especially for cases where complex topography and weather conditions are involved. This is problematic for emergency response actions, where time is of an essence. On the other hand, if the hardest part of the problem is pre-computed, the release point may be identified in only a fraction of the time required for inverse modeling. In particular, the pilot relies on historical weather data in order to create a set of climatological regimes that represent the European climate. These are subsequently employed to pre-compute the substance dispersion patterns. Then, a match between the representative atmospheric conditions and the current weather data, in line with the substance dispersion patterns, can lead to the estimation of the source location.

In the above framework, the present pilot builds upon the weather data management and searching toolkit of the 1st SC5 pilot and creates a pattern matching toolkit to facilitate the process of identifying similar weather with the support of tools aggregated on the BDE platform.

Purpose
Building on the BDE platform and on the 1st SC5 pilot, this work aims to demonstrate the following:
1. Usefulness and potential of big data technologies in nuclear emergency management.
2. An alternative methodology to the state-of-the-art inverse modelling currently used.
3. A methodology for rapid estimation of the release location of a hazardous substance.

SC5 - Climate - Pilot - Cycle 2 references	
Presentations	- https://www.big-data-europe.eu/ncsr-demokritos-presented-the-2nd-sc5-pilot-at-egu17/ - https://www.big-data-europe.eu/3rd-sc5-online-hangout/ - https://www.big-data-europe.eu/description-and-evaluation-of-1st-climate-pilot-use-case-2nd-online-hangout-wrap-up/ - https://www.big-data-europe.eu/1st-pilot-to-be-developed-in-the-frame-of-bigdataeurope-under-societal-challenge-5-climate-action-environment-resource-efficiency-and-raw-materials/ - https://www.big-data-europe.eu/sc5-bde-presentation-egu-general-assembly-vienna-17-22-april-2016/ - https://www.big-data-europe.eu/report-2nd-sc5-workshop-on-climate-action-environment-resource-efficiency-and-raw-materials/ - https://www.big-data-europe.eu/big-data-transport-workshop-presentations-are-now-online/ - https://www.big-data-europe.eu/big-data-in-the-climate-domain-online-hangout-wrap-up/
Instructions, deployment scripts and source code	https://github.com/iaklampanos/bde-climate-1

8. SC 6 - Social Sciences - pilot

The pilot is carried out by NCSR-D, and SWC in the frame of SC6 *Europe in a changing world - inclusive, innovative and reflective societies.*

The pilot is described in section 7 of deliverable D5.5.

Like SC2, the biggest challenge for this pilot is to extract the data for the various heterogeneous sources, transform it to RDF and write efficient queries that implement the required returned aggregations that are presented via a graphical web interface defined by the use cases. Here too, SWC is the responsible technical partner and uses for example their PoolParty toolkit to achieve an important part of this goal. A new feature of the second cycle, will allow us to easily add new data sources.

Furthermore, such data available in other languages will be mapped between each other to enable efficient use and comparison.

SC6 - Social Sciences - Pilot - Cycle 2 references	
Presentations	- https://www.big-data-europe.eu/bde-sc6-5th-hangout-30-05-2017-european-open-science-agenda-quo-vadis/ - https://www.big-data-europe.eu/report-sc6-4th-hangout-insight-into-virtual-currency-ecosystems/

	● https://www.big-data-europe.eu/bde-sc6-workshop-on-5-12-2016-in-cologne-report/ ● https://www.big-data-europe.eu/3258-2/ ● https://www.big-data-europe.eu/3rd-sc6-bde-hangout/ ● https://www.big-data-europe.eu/recording-of-the-sc6-bde-hangout-webinar/
Instructions, deployment scripts and source code	● https://github.com/big-data-europe/pilot-sc6-cycle1

9. SC 7 - Security - pilot

The pilot is carried out by SatCen, UoA, and NCSR-D in the frame of SC7 *Secure societies – Protecting freedom and security of Europe and its citizens*.

The pilot is described in section 8 of deliverable D5.5.

The second phase of the Societal Challenge 7 (SC7) "Secure Societies" pilot will improve and add new functionalities to the current version.

● **Change Detection**

- o *Adding further SNAP modules,* to improve the whole chain accuracy, where fundamental SNAP operators (e.g. Subset and Terrain Correction) for Sentinel 1 will be adapted to SPARK;
- o *Optimize scalability of existing Spark-based Sentinel-1 operators,* to improve the efficiency and the speed the entire workflow;
- o *Fully integrate the Pixel Aggregator module* (DBScan) into the workflow through parallelization, to improve the Result Visualization;
- o *Refinements of the output***,** to improve the classification accuracy, where the classification module will be improved to better characterize the possible changes.
- ● **Event Detection**
  - o *Dockerize SPARK version of the workflow,* to ease the deployment of the corresponding modules on different machines;
  - o *Activation of the Change Detection workflow from the Event Detection one*, to have the two workflows complementarity in the results generation;
  - o *Add the possibility to search for keywords through the Twitter Keyword Search API***,** to have a simplified search of events (without clustering and crawling of only specific sources)**.**
- ● **Whole pilot**
  - o *Cybersecurity mechanisms will be added (e.g. authentication/authorization)*, to increase the security of the platform;
  - o *Improvement of Sextant functionalities,* to improve the user experience, where current and new functionalities (derived by the second pilot phase) will be tested, based on user feedback.

SC7 - Security - Pilot - Cycle 2 references	
Presentations	- https://www.big-data-europe.eu/fourth-hangout-big-data-in-secure-societies-outcome/ - https://www.big-data-europe.eu/satcen-participated-in-the-worldcover-2017/ - https://www.big-data-europe.eu/satcen-participated-in-the-3rd-esa-international-security-symposium-iss/ - https://www.big-data-europe.eu/3rd-esa-international-security-symposium-frascati-rome-13-14-february-2017/ - https://www.big-data-europe.eu/third-hangout-big-data-in-secure-societies-outcome/ - https://www.big-data-europe.eu/bigdataeurope-second-workshop-on-big-data-in-secure-societies-outcome/ - https://www.big-data-europe.eu/second-hang-out-big-data-in-secure-societies-outcome/ - https://www.big-data-europe.eu/technical-update-on-societal-challenge-7-pilot/ - https://www.big-data-europe.eu/introducing-the-bde-societal-challenges-secure-societies/ - https://www.big-data-europe.eu/online-hang-out-big-data-in-secure-societies-outcome/
Instructions, deployment scripts	https://github.com/big-data-europe/pilot-sc7-image-aggregator

and source code	- https://github.com/big-data-europe/pilot-sc7-geotriples - https://github.com/big-data-europe/pilot-sc7-change-detector - https://github.com/big-data-europe/pilot-sc7-lookup-service

Conclusion

This report provides references to the online material that describes the implementation and deployment efforts making up this deliverable. It also provides a very brief overview of the current technical approaches used in each Societal Challenge. Managing to have an interesting pilot for each Societal Challenge being deployed on the BDE infrastructure within the provided internal deadlines was an ambitious goal which due to the good synergy between the partners was achieved satisfactory.